Text Data Augmentation for Deep Learning

Year: 2021 | Citations: 533 | Authors: Connor Shorten, T. Khoshgoftaar, B. Furht

---

Abstract

Natural Language Processing (NLP) is one of the most captivating applications of Deep Learning. In this survey, we consider how the Data Augmentation training strategy can aid in its development. We begin with the major motifs of Data Augmentation summarized into strengthening local decision boundaries, brute force training, causality and counterfactual examples, and the distinction between meaning and form. We follow these motifs with a concrete list of augmentation frameworks that have been developed for text data. Deep Learning generally struggles with the measurement of generalization and characterization of overfitting. We highlight studies that cover how augmentations can construct test sets for generalization. NLP is at an early stage in applying Data Augmentation compared to Computer Vision. We highlight the key differences and promising ideas that have yet to be tested in NLP. For the sake of practical implementation, we describe tools that facilitate Data Augmentation such as the use of consistency regularization, controllers, and offline and online augmentation pipelines, to preview a few. Finally, we discuss interesting topics around Data Augmentation in NLP such as task-specific augmentations, the use of prior knowledge in self-supervised learning versus Data Augmentation, intersections with transfer and multi-task learning, and ideas for AI-GAs (AI-Generating Algorithms). We hope this paper inspires further research interest in Text Data Augmentation.